J Inf Process Syst, Vol.21, No.2, pp.139~151, February 2025
https://doi.org/10.3745/JIPS.01.0112
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Multiple Receptive Field Anti-occlusion Self-supervision Network with CBAM for Light Field Angular Super-resolution

Wan Liao[1], Qian Zhang[1,*], Jiaqi Hou[1], Bin Wang[1], Yan Zhang[2], and Tao Yan[3]

Abstract

By recording high-dimensional light data, a light field (LF) can accurately perceive a dynamic environment, thus supporting the understanding and decision-making of intelligent systems. However, with the discrete sampling of high-dimensional signals, LF faces have insufficient efficacious acquisition of LF information. This study tackles this problem by introducing a self-supervised learning approach that uses convolutional neural networks with varying receptive fields for processing sparse view inputs and subsequently generating a dense view through warping. The primary basis relies on the fact that the inherent correlation of the LF data and the convolutional block attention module (CBAM) are applied to process the LF data and wrap the operation into a layer to construct a deep network. The proposed method eliminates occlusions and achieves super-resolution LF angle reconstruction. Extensive experiments on an HCI dataset demonstrated that the proposed model outperforms recent state-of-the-art models.

Keywords

Angular Resolution, CBAM, Light Field, Self-supervision Learning

1. Introduction

A light field (LF) is a complete representation of light flow in a 3D world. LF [1] images record both spatial and angular information and provide more comprehensive information. The entire LF can be represented by a 7D plenoptic function to simplify the function and is typically depicted by two variables. Many LF acquisition devices have been designed based on the LF plenoptic model, such as Lytro [2] and RayTrix [3]. However, the limitation of the sensor resolution leads to a trade-off between the spatial and angular resolutions. To address this issue, developing an efficient LF super-resolution, which reconstructs the original sparse LF image with a low angular resolution into a dense image with a high angular resolution, has become a key research focus.

Deep learning has various real-world applications, such as centrifugal pump fault detection [4,5]. Because of the simplicity of self-supervised learning methods, supervised learning approaches based on

Manuscript received September 28, 2022; first revision November 6, 2023; accepted November 12, 2023.

* **Corresponding Author:** Qian Zhang (qianzhang@shnu.edu.cn)
[1] School of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, China (nglih@qq.com, qianzhang@shnu.edu.cn, binwang@shnu.edu.cn)
[2] School of Computer, North China Institute of Aerospace Engineering, Langfang, China (zhangyan101@shu.edu.cn)
[3] School of Information Engineering, Putian University, Putian, China (yantaoshu@aliyun.com)

enhanced convolution techniques have been introduced [6] and extensively employed for fundamental tasks, including image classification [7]. Self-supervised methods based on deep learning are equally applicable to LF data. Most recent studies have been based on supervised learning. The Human Capital Index (HCI) [8] and Stanford Light Field datasets [9] are only two examples of real-world circumstances that may be altered; therefore, it is not necessarily practical to capture the environment in full-angle resolution. Unsupervised learning techniques have a significant value. In contrast, most current research relies on narrow baselines [10,11], whereas high spatial resolution images are frequently acquired using LF cameras with long baselines. Therefore, broad-baseline-based angle reconstruction techniques are increasingly crucial. In this study, a self-supervised convolutional neural network (CNN) method that can input broad baseline views is proposed.

Compared with the more mainstream methods proposed recently, unlike the method proposed by Jin et al. [12], our method adopts a multireceptive field structure in the network design. Unlike the method proposed by Yun et al. [13], our method considers the occlusion problem in a loss function. Our approach provides the best overall performance and speed. The significant contributions of this study are as follows.

- A multi-stream unsupervised learning network with three different sizes of receptive fields is designed for disparity estimation.
- Import the convolutional block attention module (CBAM) into the LF reconstruction network to improve network performance with fewer parameter additions.
- A novel anti-occlusion adaptive weight block matching method is proposed in the loss function design.

The remainder of this paper is organized as follows. A summary of the relevant work is provided in Section 2. Section 3 provides details on the proposed multiple-receptive anti-occlusion network using the CBAM-based angular super resolution reconstruction approach. Simulation results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related Works for Light Field Reconstruction

Some previously proposed non-learning-based approaches for LF reconstruction perform poorly in terms of speed or quality [14,15]. The two primary categories of learning-based methodologies are supervised and unsupervised/self-supervised LF reconstruction. Some methods rely on ground conditions for supervision [16], whereas others attempt to perform LF reconstruction under supervision using real depth maps [10,11,17,18]. For example, a multi-stream CNN in a feature extraction module was exploited for disparity estimation in [19]. A learning-based view synthesis method for LF cameras that considers the decomposition of disparity and color has been proposed [17]. An efficient angular super-resolution method combined with a cascaded model fusion approach was developed in [20]. Wu et al. [21] constructed an end-to-end network to reconstruct the LF. Hu et al. [22] designed a spatially angular dense network containing related blocks and spatially angular dense skip connections. In the second category of reconstruction methods, reconstruction supervision is performed by creating a variance between the reconstructed and reconstructed center maps. Several approaches [12,13,23,24] have made the depth map distort the original center view and reconstructed center map, and compensation for occlusion-based

networks has been proposed using a forward-backward warping process [24]. Li et al. [25] proposed an occlusion pattern-aware loss function for unsupervised LF disparity estimation, which successfully extracted and encoded the general occlusion patterns inherent in the LF for loss computation. Smith et al. [26] exploited a novel LF synthesizer module that reconstructs a global LF from a set of object centric LFs. Digumarti et al. [27] introduced the generalized encoding of sparse LF, allowing unsupervised learning of odometry and depth. Mousnier et al. [15] achieved high angular resolution by appropriately exploiting the image content captured at large focal length intervals. Wang et al. [28] designed an LF-InterNet to use both spatial and angular information for reconstruction.

Recently, optimized techniques have been employed for LF reconstruction. To eliminate artifacts, Wang et al. [10] used a pseudo-4DCNN paradigm. The combination of residual blocks and the CBAM has led to the development of effective network refinement structures [19]. The angular domain attention mechanism and EPI blur were implemented in a spatial-angular attention network proposed to remove the spatial high-frequency components of EPI [21,29]. Similarly, an attention-based multilevel fusion network that leads to an efficient matching cost to eliminate occlusions was proposed [30].

The performance of LF reconstruction was improved using deep-learning-based approaches. Most deep-learning-based approaches, however, do not consider various receptive fields for the input of the sparse view, which is also important for enhancing the quality of LF reconstruction. Additionally, there is still some space to remove occlusion-related artifacts and calculate a densely rebuilt LF using a new loss function.

This paper addresses these challenges by proposing a self-supervised learning technique that produces dense view disparity maps by feeding input from sparse views into multiple CNNs with various receptive fields, and then obtains dense views through depth image-based rendering. [31] served as an inspiration for CBAM, which efficiently extracted detailed information by fusing spatial and channel attention. Examples of CBAM applications in object recognition include YOLOv4 [32], super-resolution [33,34], and pose estimation [35]. Therefore, we used the CBAM to discover strong correlations between related LF perspectives. To recover details more accurately, we defined a unique loss function. The results of our experiments demonstrate the effectiveness of our strategy in producing unique vistas while enhancing the realistic features.

3. Proposed LF Angular Super-resolution Reconstruction Framework

Fig. 1 depicts our proposed LF angular super-resolution reconstruction framework. It has three primary modules:

1) The multi-stream disparity estimation module (MDEM): predicts an advanced disparity map for each LF input view.

2) The LF warping module (LFWM): creates new views by adjusting the input views using the estimated disparity map.

3) The LF blending module (LFBM): eliminates artifacts and enhances network performance, producing the final high angular resolution LF.

Fig. 1 shows the reconstruction of a 3×3 LF from a 2×2 sparse LF. Reconstruction at other angular resolutions can be easily achieved by tuning certain network architecture parameters.

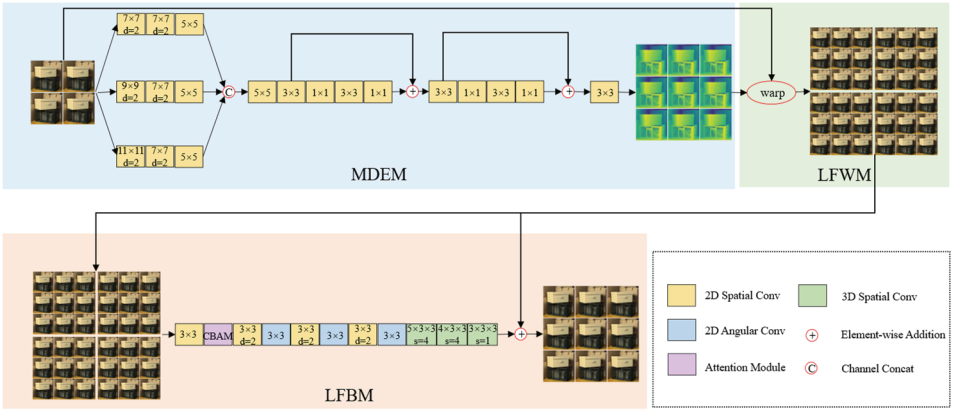


Fig. 1. Multi-stream attention fusion CNN architecture diagram. MDEM is a multi-receptive field network with residual structure, and LFBM is a single-stream network with an attention mechanism.

3.1 Multi-stream Parallax Estimation Module

First, a low-resolution input view is fed into a three-stream CNN with different receptive fields. The multi-stream structure contains a dilated convolution with a convolution kernel size of 7×7, 9×9, or 11×11, a dilated convolution with the same convolution kernel size of 7×7, and a normal convolution with a kernel size of 5×5. The three channels were then merged, delivered to two CNNs with residual blocks, and finally passed through a 3×3 convolution kernel to obtain the final disparity maps (taking 2×2 to 3×3 as an example). In this module, this process can be expressed as follows:

$$D(uu,x) = f_p\big(L(u,x)\big), \tag{1}$$

where f_p represents MDEM and uu refers to the angular resolution of the disparity map. To distinguish this map from the original map u, we introduce a uu representation.

3.2 Light Field Warping Module

The warping module uses the four input views and nine disparity map outputs of the multi-stream parallax estimation module to perform physical distortion operations and obtains 4×9, a total of 36 sub-images. The warped view can be expressed as follows:

$$W(u,uu,x) = f_w(L(u,x), D(uu,x)), \tag{2}$$

f_w refers to the warp operation, u represents warping by the u' input view and disparity map, uu represents warping by the uu' disparity map and input view, and x represents the spatial resolution.

Knowing that disparity refers to the offset of pixels, all that is required to create the final warp map is to multiply the disparity map of the relevant point by the baseline length and then add the resulting offset to the original view. The resulting parallax increases with the distance from the central vision and is mathematically stated as follows:

$$W(u',uu',x) = L\big(u', x + D(uu',x)(uuc - u')\big), \tag{3}$$

where $D(uu',x)$ is the disparity map obtained by the CNN in this study; the disparity map of the real scene is not used.

3.3 Light Field Blending Module

After finishing the warp module, some researchers completed LF reconstruction; however, the experimental data revealed that these techniques have numerous problems and operate poorly. To enhance experimental performance, we included an LF blending module.

First, the 4×9 warped view was mixed into an ordinary 2D spatial convolution. Subsequently, feature maps are sent into CBAM, followed by an expanded convolution with a kernel dilation rate of 2, and then go through a 2D angular convolution, the size of which is 3×3. Repeat these 2 alternates four times. Then, the network terminates with three 3D convolution layers; the convolution kernel sizes are 5×3×3, 4×3×3, and 3×3×3, and the step sizes are 4×1×1, 4×1×1, and 1×1×1. To acquire the final reconstruction result map, the final output feature maps were allowed to have the same number of channels as in the warping view. The above process can be expressed as:

$$\hat{L}(u,x) = W(u,uu,x) + f_b\big(W(u,uu,x)\big), \tag{4}$$

where f_b represents the LF blending network.

3.4 Convolutional Block Attention Module

The CBAM [31] is used to adaptively improve features by successively inferring the attention map along the channel and spatial order and then multiplying the attention map by the input feature map. The network architecture is shown in Fig. 2.

The entire process is outlined as follows:

$$F' = \sigma\big(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))\big) \otimes F, \tag{5}$$

$$F'' = \sigma\big(f([\text{AvgPool}\,(F'); \text{MaxPool}\,(F')])\big) \otimes F'. \tag{6}$$

where σ represents the sigmoid function and f represents the convolution operation, MLP represents shared parameters convolution operation, $\otimes$ represents element-wise multiplication.

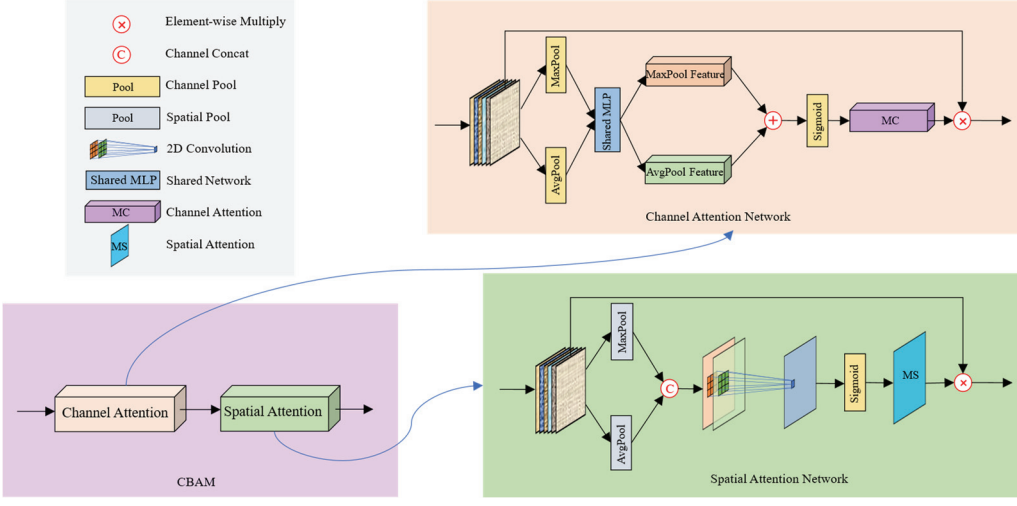


Fig. 2. CBAM overview. The first half is the channel attention module, and the second half is the spatial attention module.

3.5 Training Loss

Some existing methods based on unsupervised learning all use unilateral and limited loss functions, which typically offer some artifacts to the reconstructed LF. To mitigate this, in this study, we design six loss functions to constrain the network, which are expressed as follows:

Map loss: To create a mapping connection, the gradient of the disparity map must be used to balance the difference between the input and warped views. This is known as map loss, and is expressed as follows:

$$\ell_{\text{map}} = \sum_{x,u}\left(\sum_{uu}|L(u,x) - W(u,uu,x)| + \nabla_x D(u,x)\right), \tag{7}$$

where u, uu represents the angular resolution, x represents the spatial resolution, L represents the input LF view. W represents the view after warp, $\nabla_x D$ represents the gradient of disparity.

EPI gradient loss: The EPI is created as a 2D slice by dropping the spatial and angular capabilities of the LF in 1D. We set the input and output view EPI gradients to be equal because parallax has a direct relationship with the EPI image slope magnitude. The loss function is expressed as follows:

$$\begin{aligned}\ell_{epi} = \sum_{x,u}\left(\left|\nabla_y E_{x,u}(v,y) - \nabla_y \hat{E}_{x,u}(v,y)\right| + \left|\nabla_v E_{x,u}(v,y) - \nabla_v \hat{E}_{x,u}(v,y)\right|\right) \\ + \sum_{y,v}\left(\left|\nabla_x E_{y,v}(u,x) - \nabla_x \hat{E}_{y,v}(u,x)\right| + \left|\nabla_u E_{y,v}(u,x) - \nabla_u \hat{E}_{y,v}(u,x)\right|\right)\end{aligned} \tag{8}$$

Blend loss: Because we used the corner view to reconstruct the middle view, we made the reconstructed corner view equal to the original input corner view. This loss occupies a relatively large weight in the total loss function and is expressed as follows:

$$\ell_{blend} = \sum_{x,u}\left|\hat{L}(u_{\text{corner}},x) - L(u_{\text{corner}},x)\right|, \tag{9}$$

where $\hat{L}$ represents the output LF view, u_{corner} represents the view from the four corners.

Warp loss: After the physical warping procedure, nine reconstructed views were obtained for a single input view, with the center view of each view remaining consistent with the corresponding original input view. These are summarized as follows:

$$\ell_{\text{warp}} = \sum_{x}\left(\sum_{i=1}^{in^2}|L(u_i,x) - W(u_i,uu_{\text{center}},x)|\right). \tag{10}$$

Anti-occlusion loss: This is one of the highlights of the present study. The AM was used for the second time. This loss function is primarily used to eliminate the impact of occlusion and is expressed as follows:

$$\ell_{\text{anti-occlusion}} = \sum_{i=1}^{out^2}\sum_{x}\left|\hat{L}(u_i,x) - \sum_{\text{j=1}}^{in^2}\omega_{i,j}W\left(u_j,uu_i,x\right)\right|, \tag{11}$$

where out^2 represents the number of final reconstructed output views, in^2 represents the number of input warp views, where $\omega_{i,j}$ represents the weight of different input views, the farther away from the output view, the greater the weight, and the more it can get around obstacles.

Parallax loss: Because the baseline of the LF camera is short, the disparity maps produced at various optical center positions are rather small, and the distance between adjacent disparity maps is even smaller, allowing us to make the intermediate disparity maps as similar as possible. Otherwise, it occupies the

leading position. The loss function is expressed as follows:

$$\ell_{parallax} = \sum_{i=1}^{out^2} \sum_{x} |D(u_{center}, x) - D(u_i, x)|. \quad (12)$$

Total loss: Our total loss function is defined as follows:

$$L_{total} = \alpha \ell_{map} + \beta \ell_{epi} + \gamma \ell_{blend} + \delta \ell_{warp} + \varepsilon \ell_{anti-occlusion} + \xi \ell_{parallax}, \quad (12)$$

where α, β, γ, δ, ε, ξ are constants, and we set to 4, 8, 2, 2, 2, 1 in our experiment.

4. Experiments

Quantitative and qualitative comparison studies have been conducted using various cutting-edge methods, including those developed by Jin et al. [12], Yun et al. [13], Wu et al. [36], and Meng et al. [37]. To achieve fairness, we first trained all the experiments using the settings suggested by the authors and then tested them using the same hardware and environment on the new HCI LF dataset.

The HCI dataset contains 28 scenes, including four training scenes, four test scenes, 16 additional scenes, and four stratified scenes. During training, four training scenes and 15 additional scenes except "dishes" are selected, 19 scenes in total as the training set. During the test, three test scenes containing "bedroom," "bicycle," "herb" and an additional scene containing "dishes," a total of four scenes are selected as the test sets. The single scene size of the HCI dataset was 512×512×9×9; we reconstructed the input view from 2×2 to 7×7, and the reconstructed result had a size of 468×468×7×7; thus, when comparing.

4.1 Quantitative Assessment

The peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are the two key metrics employed for the quantitative analysis of LF reconstruction in our study. The experimental results are presented in Table 1. As is evident from the data, the techniques proposed by Wu et al. [36] and Meng et al. [37] exhibit strong performance under straightforward conditions, however, face challenges in obstructed scenes. By contrast, our approach demonstrates remarkable robustness across diverse scenarios, consistently outperforming alternative methods. Moreover, the graphical representation of the running duration and reconstruction performance, as depicted in Fig. 3, underscores the exceptional speed and efficacy of our method.

The practical implications of these quantitative results extend beyond statistical analysis. Our method's consistent excellence in challenging scenarios, where others falter, underscores its potential significance in fields such as autonomous navigation, 3D imaging, virtual reality (VR), and various related domains.

Table 1. Quantitative comparisons (PSNR/SSIM) of different methods

	Jin et al. [12]	Yun et al. [13]	Wu et al. [36]	Meng et al. [37]	This work
bedroom	37.97/0.951	37.58/0.950	39.12/0.967	38.68/0.958	**39.66/0.973**
bicycle	31.22/0.919	30.86/0.916	30.80/0.914	31.12/0.926	**31.58/0.927**
herbs	31.97/0.857	32.20/0.866	30.77/0.838	31.09/0.848	**32.45/0.868**
dishes	28.36/0.910	28.34/0.909	26.56/0.891	28.46/**0.915**	**28.62**/0.914
Avg	32.38/0.909	32.25/0.910	31.81/0.903	32.34/0.912	**33.08/0.921**

The bold font indicates the best performance in each test.

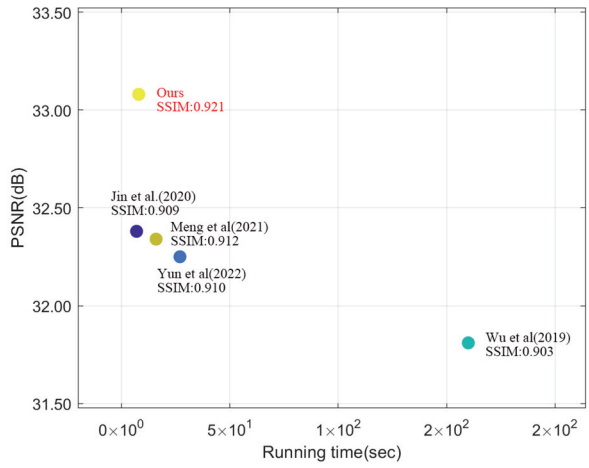


Fig. 3. Comparison of LF reconstruction effects achieved by different methods in recent years. Comparisons were made from reconstruction quality (PSNR, SSIM, and runtime).

4.2 Qualitative Analysis

In Fig. 4, the residual plot uses the center view as a baseline, a local close-up in the red box, and normalized red and blue levels to depict the difference. When compared vertically, we can see that our method is better than the first four methods in each scene, our technique has less incorrect pixels and only a few edges part have large differences, and the rest are relatively smooth. Additionally, our test performs well, particularly in scenarios containing occlusions.

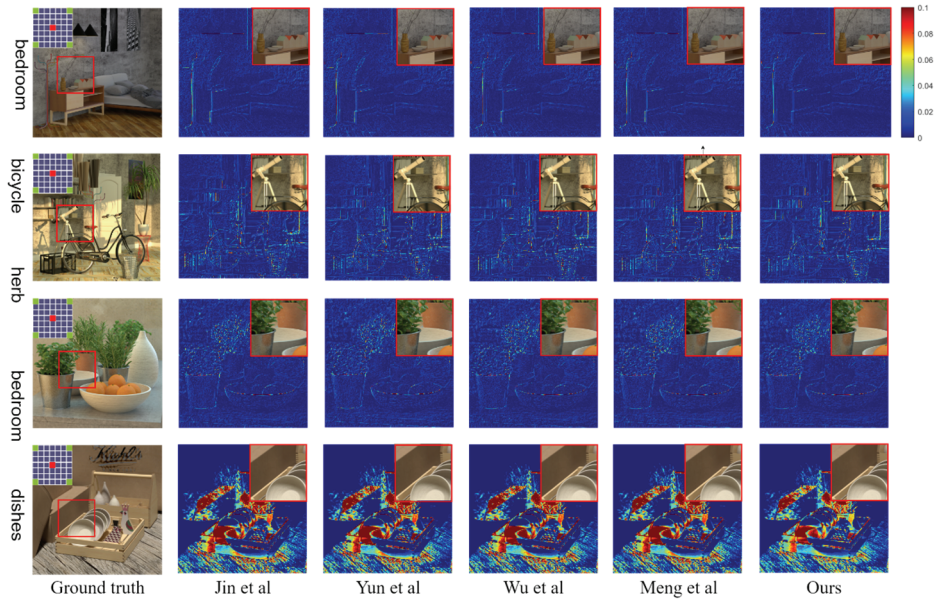


Fig. 4. Comparison of our method with other methods. The results show the ground-truth image, the residual plot of the synthesized center view versus the ground-truth image, a close-up version of the portion of the image boxed in red. In the grid in the upper left corner of each ground truth image, the four-corner blue box represents the input view in the LF, and the red box represents the display view.

4.3 Ablation Study

Effects of multi-stream network: We compare the impact of multi-stream networks on the reconstruction performance in Table 2. The parameter settings for each network model are listed in Table 2. The multi-stream LF blending module has three-stream networks with varying convolution kernel sizes, which are then merged in a fair proportion. The single-stream network employed a 9×9 receptive field for the disparity estimation. According to ablation studies, the multi-receptive field used by the parallax estimation module can significantly improve the network performance.

Effects of CBAM: The impact of a network's CBAM structure on the network performance in various scenarios is shown in Table 3. A 5×5 convolution kernel was placed in front of the two CBAMs of the disparity estimation module, and a 3×3 convolution operation was placed after the warp view and first residual block following the multi-stream network connection. A comparison of the four network architectural parameters is presented in Table 3. According to the experimental data, adding a CBAM module to the LF hybrid module can vastly improve the network performance while only adding a few extra parameters.

Table 2. Comparison of PSNR/SSIM with(+) or without(-) multi-stream structure in MDEN and LFBN

	MDEN- LFBN-	MDEN+ LFBN-	MDEN- LFBN+	MDEN+ LFBN+
bedroom	37.98/0.952	**38.77/0.962**	37.68/0.949	38.01/0.956
bicycle	31.21/0.919	**31.56/0.921**	30.87/0.914	31.12/0.917
herbs	31.98/0.858	**32.08/0.865**	31.34/0.842	32.05/0.863
dishes	28.37/0.910	**28.82/0.917**	28.02/0.904	28.73/0.912
Avg	32.39/0.910	**32.81/0.916**	31.98/0.902	32.48/0.912
Parameters	**11.91M**	19.34M	12.13M	19.56M

The bold font indicates the best performance in each test.

Table 3. Comparison of PSNR/SSIM with(+) or without(-) CBAM in MDEN and LFBN

	MDEN- LFBN-	MDEN++ LFBN-	MDEN- LFBN+	MDEN++ LFBN+
bedroom	38.77/0.962	38.66/0.959	**39.66/0.973**	39.12/0.966
bicycle	31.56/0.921	31.23/0.911	**31.58/0.927**	31.42/0.912
herbs	32.08/0.865	31.98/0.859	**32.45/0.868**	32.15/0.866
dishes	**28.82/0.917**	28.76/0.916	28.62/0.914	28.60/0.914
Avg	32.81/0.916	32.66/0.911	**33.08/0.921**	32.82/0.915
Parameters	**19.34M**	19.36M	19.36M	19.38M

The bold font indicates the best performance in each test.

5. Conclusion

In this study, we introduce a CNN architecture featuring multiple receptive fields and a CBAM structure for LF angular super-resolution reconstruction. This approach leverages input views to extract richer data using features from various receptive fields. Although our model exhibits a high adaptability and can be trained without relying on real depth maps or scene graphs, we must acknowledge its limitations. The following aspects should be addressed in future research:

1) Model limitations: Our model has certain limitations. For instance, it may face challenges in

scenarios with extreme variations in lighting conditions. Furthermore, the computational cost, particularly in terms of the model parameters, remains a concern.

2) Improvement suggestions: To enhance the model performance, future studies should focus on reducing its computational overhead and improving its robustness to diverse lighting conditions. In addition, further exploration of loss-function variations and potential regularization techniques should be considered to effectively mitigate occlusion-related artifacts.

3) Future directions: This study opens new avenues for future research. These include investigating the integration of real-time depth map estimation and scene graph generation to further enhance the performance and versatility of the model. The exploration of lightweight model architectures and their effect on efficiency is also a potential direction.

In summary, the proposed model demonstrated competitive performance, as evidenced by extensive experiments on the HCI dataset. However, we must address these limitations, implement the suggested improvements, and explore the suggested future directions for advancing the field of LF angular super-resolution reconstruction.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This study was jointly supported by the National Natural Science Foundation of China (Grant No. 62301320), Natural Science Foundation of Fujian (No. 2023J011009), Scientific Research Project of Putian Science and Technology Bureau (No. 2021G2001ptxy08), and the Colleges and Universities in Hebei Province Science and Technology Research Project (No. ZC2021006).

References

[1] A. Gershun, “The light field,” *Journal of Mathematics and Physics*, vol. 18, no. 1-4, pp. 51-151, 1939. https://doi.org/10.1002/sapm193918151

[2] The lytro website [Online]. Available: https://www.lytro.com/.

[3] The raytrix website [Online]. Available: http://www.raytrix.de/.

[4] S. Ahmad, Z. Ahmad, and J. M. Kim, “A centrifugal pump fault diagnosis framework based on supervised contrastive learning,” *Sensors*, vol. 22, no. 17, article no. 6448, 2022. https://doi.org/10.3390/s22176448

[5] A. E. Prosvirin, Z. Ahmad, and J. M. Kim, “Global and local feature extraction using a convolutional autoencoder and neural networks for diagnosing centrifugal pump mechanical faults,” *IEEE Access*, vol. 9, pp. 65838-65854, 2021. https://doi.org/10.1109/ACCESS.2021.3076571

[6] Y. Oh, M. Jeon, D. Ko, and H. J. Kim, “Randomly shuffled convolution for self-supervised representation learning,” *Information Sciences*, vol. 623, pp. 206-219, 2023. https://doi.org/10.1016/j.ins.2022.11.022

[7] K. Liu, R. Meng, L. Li, J. Mao, and H. Chen, “SiSL-Net: saliency-guided self-supervised learning network for image classification,” *Neurocomputing*, vol. 510, pp. 193-202, 2022. https://doi.org/10.1016/j.neucom.

2022.09.029

[8] S. Wanner, S. Meister, and B. Goldluecke, "Datasets and benchmarks for densely sampled 4D light fields," *VMV Vision, Modeling, and Visualization*, vol. 13, pp. 225-226, 2013. https://doi.org/10.2312/PE.VMV.VMV13.225-226

[9] Stanford Graphics Laboratory, "The (new) Stanford Light Field Archive," 2008 [Online]. Available: http://lightfield.stanford.edu/.

[10] Y. Wang, F. Liu, Z. Wang, G. Hou, Z. Sun, and T. Tan, "End-to-end view synthesis for light field imaging with pseudo 4DCNN," in *Computer Vision - ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 340-355. https://doi.org/10.1007/978-3-030-01216-8_21

[11] G. Wu, Y. Liu, L. Fang, Q. Dai, and T. Chai, "Light field reconstruction using convolutional network on EPI and extended applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1681-1694, 2019. https://doi.org/10.1109/TPAMI.2018.2845393

[12] J. Jin, J. Hou, H. Yuan, and S. Kwong, "Learning light field angular super-resolution via a geometry-aware network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11141-11148, 2020. https://doi.org/10.1609/aaai.v34i07.6771

[13] S. Yun, J. Jang, and J. Paik, "Geometry-aware light field angular super resolution using multiple receptive field network," in *Proceedings of 2022 International Conference on Electronics, Information, and Communication (ICEIC)*, Jeju, South Korea, 2022, pp. 1-3. https://doi.org/10.1109/ICEIC54506.2022.9748458

[14] S. Wanner and B. Goldluecke, "Variational light field analysis for disparity estimation and super-resolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 606-619, 2019. https://doi.org/10.1109/TPAMI.2013.147

[15] A. Mousnier, E. Vural, and C. Guillemot, "Partial light field tomographic reconstruction from a fixed-camera focal stack," 2015 [Online]. Available: https://arxiv.org/abs/1503.01903.

[16] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "DeepStereo: learning to predict new views from the world's imagery," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 5515-5524. https://doi.org/10.1109/CVPR.2016.595

[17] N. K. Kalantari, T. C. Wang, and R. Ramamoorthi, "Learning-based view synthesis for light field cameras," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 6, article no. 193, 2016. https://doi.org/10.1145/2980179.2980251

[18] H. W. F. Yeung, J. Hou, J. Chen, Y. Y. Chung, and X. Chen, "Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues," in *Computer Vision – ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 138-154. https://doi.org/10.1007/978-3-030-01231-1_9

[19] M. S. K. Gul, M. U. Mukati, M. Batz, S. Forchhammer, and J. Keinert, "Light-field view synthesis using a convolutional block attention module," in *Proceedings of 2021 IEEE International Conference on Image Processing (ICIP)*, Anchorage, AK, USA, 2021, pp. 3398-3402. https://doi.org/10.1109/ICIP42928.2021.9506586

[20] F. Cao, P. An, X. Huang, C. Yang, and Q. Wu, "Multi-models fusion for light field angular super-resolution," in *Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021, pp. 2365-2369. https://doi.org/10.1109/ICASSP39728.2021.9413824

[21] G. Wu, Y. Wang, Y. Liu, L. Fang, and T. Chai, "Spatial-angular attention network for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 30, pp. 8999-9013, 2021. https://doi.org/10.1109/TIP.2021.3122089

[22] Z. Hu, H. W. F. Yeung, X. Chen, Y. Y. Chung, and H. Li, "Efficient light field reconstruction via spatio-angular dense network," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, article no. 5014314, 2021. https://doi.org/10.1109/TIM.2021.3100326

[23] J. Peng, Z. Xiong, D. Liu, and X. Chen, "Unsupervised depth estimation from light field using a convolutional neural network," in *Proceedings of 2018 International Conference on 3D Vision (3DV)*, Verona, Italy, 2018, pp. 295-303. https://doi.org/10.1109/3DV.2018.00042

[24] L. Ni, H. Jiang, J. Cai, J. Zheng, H. Li, and X. Liu, "Unsupervised dense light field reconstruction with occlusion awareness," *Computer Graphics Forum*, vol. 38, no. 7, pp. 425-436, 2019. https://doi.org/10.1111/cgf.13849

[25] P. Li, J. Zhao, J. Wu, C. Deng, Y. Han, H. Wang, and T. Yu, "OPAL: occlusion pattern aware loss for unsupervised light field disparity estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 2, pp. 681-694, 2024. https://doi.org/10.1109/TPAMI.2023.3296600

[26] C. Smith, H. X. Yu, S. Zakharov, F. Durand, J. B. Tenenbaum, J. Wu, and V. Sitzmann, "Unsupervised discovery and composition of object light fields," 2022 [Online]. Available: https://arxiv.org/abs/2205.03923.

[27] S. T. Digumarti, J. Daniel, A. Ravendran, R. Griffiths, and D. G. Dansereau, "Unsupervised learning of depth estimation and visual odometry for sparse light field cameras," in *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague, Czech Republic, 2021, pp. 278-285. https://doi.org/10.1109/IROS51168.2021.9636570

[28] Y. Wang, L. Wang, J. Yang, W. An, J. Yu, and Y. Guo, "Spatial-angular interaction for light field image super-resolution," in *Computer Vision–ECCV 2020*. Cham, Switzerland: Springer, 2020, pp. 290-308. https://doi.org/10.1007/978-3-030-58592-1_18

[29] G. Wu, M. Zhao, L. Wang, Q. Dai, T. Chai, and Y. Liu, "Light field reconstruction using deep convolutional network on EPI," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 1638-1646. https://doi.org/10.1109/CVPR.2017.178

[30] J. Chen, S. Zhang, and Y. Lin, "Attention-based multi-level fusion network for light field depth estimation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, pp. 1009-1017, 2021. https://doi.org/10.1609/aaai.v35i2.16185

[31] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*. Cham, Switzerland: Springer, 2018, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[32] A. Bochkovskiy, C. Y. Wang, and H. Y. M. Liao, "YOLOv4: optimal speed and accuracy of object detection," 2020 [Online]. Available: https://arxiv.org/abs/2004.10934.

[33] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. Change Loy, "EDVR: video restoration with enhanced deformable convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1954-1963. https://doi.org/10.1109/CVPRW.2019.00247

[34] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice, France, 2019, pp. 2024-2032. https://doi.org/10.1145/3343031.3351084

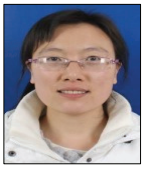
[35] T. Y. Yang, Y. T. Chen, Y. Y. Lin, and Y. Y. Chuang, "FSA-Net: learning fine-grained structure aggregation for head pose estimation from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 1087-1096. https://doi.org/10.1109/CVPR.2019.00118

[36] G. Wu, Y. Liu, Q. Dai, and T. Chai, "Learning sheared EPI structure for light field reconstruction," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3261-3273, 2019. https://doi.org/10.1109/TIP.2019.2895463

[37] N. Meng, X. Wu, J. Liu, and E. Lam, "High-order residual network for light field super-resolution," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11757-11764, 2020. https://doi.org/10.1609/aaai.v34i07.6847

Wan Liao **https://orcid.org/0000-0002-1300-3580**

He graduated from the School of Information and Mechatronics Engineering of Shanghai Normal University with a bachelor's degree in communications. Currently studying for a master's degree in electronic information at Shanghai Normal University. His research interests include light field depth estimation and angle reconstruction.

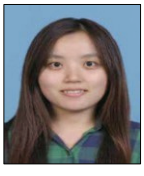


Qian Zhang **https://orcid.org/0000-0003-0760-9241**

She is now the associate professor of Shanghai normal University, China. She received her Ph.D. from Shanghai University in China. Her research interest fields include video processing.

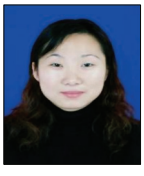
Jiaqi Hou **https://orcid.org/0000-0002-7363-7753**

She received her bachelor’s degree in electronic science and technology from the College of Electronic Engineering, Heilongjiang University. She is currently studying for a master’s degree in electronic information at Shanghai Normal University. Her research interest includes super-resolution reconstruction of light field images.



Bin Wang **https://orcid.org/0000-0002-5860-3440**

She received her Ph.D. degree from Department of Automation, Shanghai Jiao Tong University, Shanghai, China. She is now the associate professor of Shanghai Normal University, China. Her research interests include computer vision, machine learning, image processing, and multimedia analysis.



Yan Zhang **https://orcid.org/0000-0001-5970-7244**

She received the Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China. She has been with the faculty of the School of Computer, North China Institute of Aerospace Engineering, where she is currently an associate professor. Her major research interests include Stereo video quality assessment, artificial intelligence.

Tao Yan **https://orcid.org/0000-0002-8304-8733**

He received the Ph.D. degrees in communication and information systems from Shanghai University, Shanghai, China, in 2010. He has been with the faculty of the School of Information Engineering, Putian University, where he is currently an associate professor. His major research interests include multiview high efficiency video coding, rate control, and video codec optimization. He has authored or co-authored more than 20 refereed technical papers in international journals and conferences in the field of video coding and image processing. He currently presides National Natural Science Foundation project.